SONiCS: PCR stutter noise correction in genome-scale microsatellites

Kedzierska, Katarzyna Z ; Gerber, Livia ; Cagnazzi, Daniele ; Krützen, Michael ; Ratan, Aakrosh ; Kistler, Logan

Abstract: Motivation Massively parallel capture of short tandem repeats (STRs, or microsatellites) provides a strategy for population genomic and demographic analyses at high resolution with or without a reference genome. However, the high Polymerase Chain Reaction (PCR) cycle numbers needed for target capture experiments create genotyping noise through polymerase slippage known as PCR stutter. Results We developed SONiCS—Stutter mONte Carlo Simulation—a solution for stutter correction based on dense forward simulations of PCR and capture experimental conditions. To test SONiCS, we genotyped a 2499-marker STR panel in 22 humpback dolphins (Sousa sahulensis) using target capture, and generated capillary-based genotypes to validate five of these markers. In these 110 comparisons, SONiCS showed a 99.1% accuracy rate and a 98.2% genotyping success rate, miscalling a single allele in a marker with low sequence coverage and rejecting another as un-callable. Availability and implementation Source code and documentation for SONiCS is freely available at https://github.com/kzkedzierska/sonics. Raw read data used in experimental validation of SONiCS have been deposited in the Sequence Read Archive under accession number SRP135756.

DOI: https://doi.org/10.1093/bioinformatics/bty485

SONiCS: PCR stutter noise correction in genome-scale microsatellites

Katarzyna Z. Kedzierska[1], Livia Gerber[2], Daniele Cagnazzi[3], Michael Krützen[2], Aakrosh Ratan*[§1], Logan Kistler*[§4]

[1] Department of Public Health Sciences and Center for Public Health Genomics, University of Virginia, Charlottesville 22908, VA, USA

[2] Evolutionary Genetics Group, Department of Anthropology, University of Zurich, CH-8057 Zurich, Switzerland.

[3] School of Environment Science and Engineering, Marine Ecology Research Centre, Southern Cross University, Lismore 2480, Australia.

[4] Department of Anthropology, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560 USA

§These authors contributed equally to the work.

Abstract

Summary
Massively parallel capture of short tandem repeats (STRs, or microsatellites) provides a strategy for population genomic and demographic analyses at high resolution with or without a reference genome. However, the high Polymerase Chain Reaction (PCR) cycle numbers needed for target capture experiments create genotyping noise through polymerase slippage known as PCR stutter. We developed SONiCS—Stutter mONte Carlo Simulation—a solution for stutter correction based on dense forward simulations of PCR and capture experimental conditions. To test SONiCS, we genotyped a 2499-marker STR panel in 22 humpback dolphins (*Sousa sahulensis*) using target capture, and generated capillary-based genotypes to validate 5 of these markers. In these 110 comparisons, SONiCS showed a 99.1% accuracy rate and a 98.2% genotyping success rate, miscalling a single allele in a marker with low sequence coverage and rejecting another as un-callable.

Availability and Implementation
Source code and documentation for SONiCS is freely available at https://github.com/kzkedzierska/sonics. Raw read data used in experimental validation of SONiCS have been deposited in the Sequence Read Archive under accession number SRP135756. Additional details are provided in Supplementary Online Material (SOM).

Contact
Logan Kistler – KistlerL@si.edu; +1 202 633 1908
Aakrosh Ratan – ratan@virginia.edu; +1 434 982 6583

Main Text

1 Introduction

For species lacking reference genomes, reduced representation sequencing methods such as target capture (Gnirke *et al.*, 2009), RAD-seq (Baird *et al.*, 2008), and genotyping-by-sequencing (Elshire *et al.*, 2011) can yield efficient datasets suitable for a wide range of genomic applications. These methods are typically used to develop sets of single nucleotide polymorphisms (SNPs), but target capture of genome-wide short tandem repeats (STRs) has recently emerged as a strategy for generating massively parallel datasets with rapid rates of evolution (Kistler *et al.*, 2017). STR capture circumvents the workflow bottlenecks of traditional STR development and genotyping while enhancing resolution in population genomic contexts compared with strictly SNP-based approaches. Furthermore, analyses of linked and co-phased SNPs and STRs allows for control of homoplasy that is frequently observed at STR loci, which can be identical by state without being identical by descent due to their high mutability (Ellegren, 2004).

A key challenge for STR capture is PCR stutter—the physical slippage of DNA polymerase on the template strand causing molecules with different number of repeats of the motif sequence to be synthesized (Schlötterer and Tautz, 1992). Stutter is a well-known obstacle for traditional STR genotyping based on amplicon size, and it remains confounding when using genomic STR methods (Gymrek *et al.*, 2012; Kistler *et al.*, 2017). In target capture experiments, a library amplification step is used both before and after probe–library hybridization totaling 20 or more PCR cycles (SOM), increasing opportunities for stutter-based alleles to appear and propagate through successive PCR steps. To address this issue, we developed a method for fitting the best diploid STR genotype to a set of raw allele counts by comparing the results of dense forward simulations.

2 Methods

The complete SONiCS method is described in SOM, and outlined in Figure 1. Briefly, the

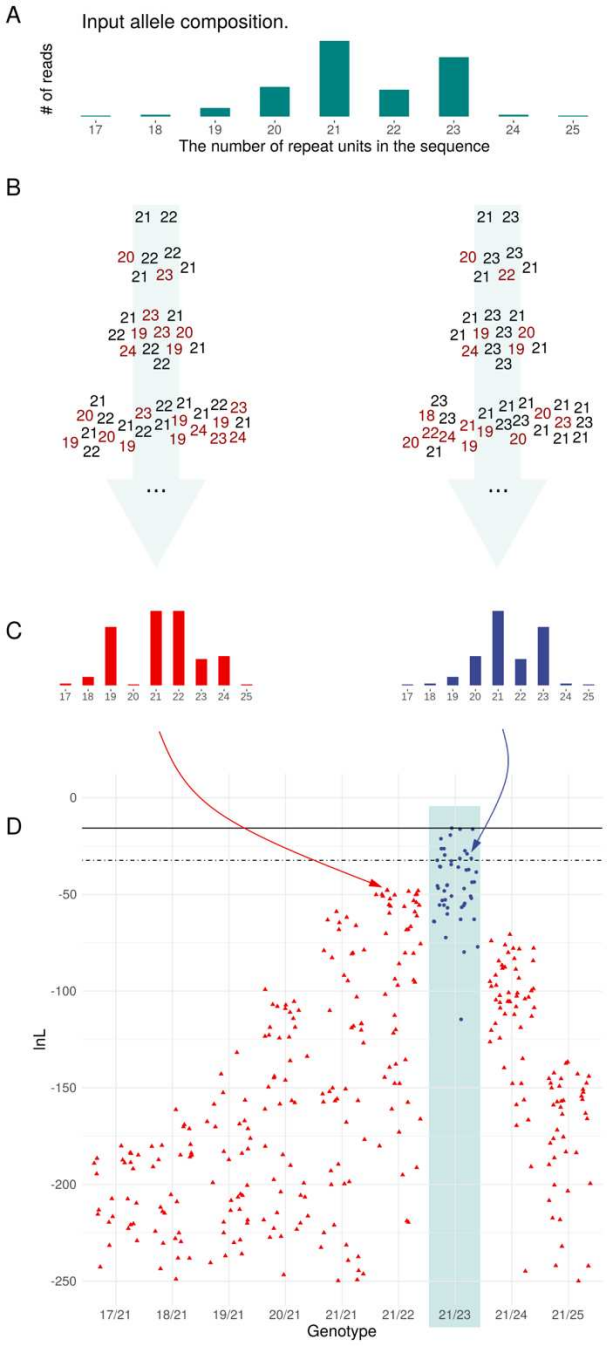

Figure 1 – Schematic of the method starting from raw read support to a genotype call. The user provides an empirical set of raw reads supporting alleles with different repeat counts (A). Several reactions that include PCR and the capture step are modeled in-silico starting from various combination of input alleles, creating a post-PCR pool of molecules (C; new alleles in each cycle are shown in red). The log-likelihood (lnL) of sampling the starting read set from the output PCR pool is calculated as descried in SOM, along with the coefficient of determination and the percent identity for alternative filtering schemes. Next, the distributions of lnL values between genotypes (D) are compared using the Mann-Whitney U test and lnL ratios are calculated. A genotype is called if conditions of both tests are met. In this example, the input raw read support consists of alleles ranging from 17 to 25 repeat units, and SONiCS was able to determine the genotype of this particular loci as 21/23 after 500 simulations. Solid and dashed lines show the best single lnL and 75th percentile lnL under the best supported genotype.

user provides a set of raw reads supporting an unknown genotype, and a starting pool of simulated molecules is generated using two independent alleles selected from the set of all possible genotypes present. A set of reaction parameters including the efficiency of amplification, efficiency of capture, and the probability of polymerase slippage modeled separately for insertions and deletions are then drawn from weak uniform priors based on experimental observations or user inputs. The complete PCR and capture process is then modeled in-silico under the chosen parameters, and we calculate the likelihood of observing the input dataset from the product of the simulated starting genotype. After running a large number of independent simulations, we compare the distributions of the log likelihoods between all possible pairs of genotype calls that could be made. We then call a genotype on basis of a Bonferroni-corrected Mann Whitney U test p-value and the likelihood ratios between the best fitting and second best fitting genotype.

SONiCS accepts either a single genotype provided at the command line or a VCF file including ALLREADS, MOTIF, and REF fields as produced by *allelotype,* a component of lobSTR (Gymrek *et al.*, 2012). SONiCS writes a summary genotype file including the Mann Whitney U test results, likelihood ratios for best and third quartile lnL values between alternative genotypes, and the number of trials conducted. Optionally, SONiCS can also report the verbose parameters and results of each individual simulation for under all tested genotypes. This functionality allows completely flexible interrogation of simulation results and alternative filtering schemes for genotype selection according to the specific needs of stringency and experimental design. In order to optimize performance, large parts of the code for SONiCS are written in Cython (Behnel *et al.*, 2011), a superset of the Python programming language designed to give C-like performance with code mostly written in Python. The software is capable of using multiple processors and uses the ‘multiprocessing’ package from Python. On a single processor, SONiCS calculates 6-9 genotypes per minute on average. The signal to noise ratio at an STR locus depends on the coverage at the locus, the extent of the stutter, the repeat motif, the

distribution of biological alleles, and several other factors. Based on sub-sampling at validated loci (Fig S5, Supplementary Methods), we recommend using SONiCS on loci with a minimal coverage of 45 reads as a conservative threshold.

To test the accuracy of SONiCS, we used BaitSTR (Kistler *et al.*, 2017) and targeted resequencing to generate a set of 2499 STRs in a set of 22 Australian humpback dolphins (*Sousa sahulensis*; target regions provided as Supplemental Dataset S1). We then used lobSTR (Gymrek *et al.*, 2012) to align reads and summarize raw read support for target STRs and SONiCS to calculate resulting STR genotypes. We used traditional PCR and capillary genotyping to analyze five of the captured STR loci as a truth-set for comparison with SONiCS results (Table S1). Complete genotyping procedures are described in SOM. Tissue samples were originally collected under permits from the Queensland Department of Environment and Heritage protection (WISP16457615) and combined permit of the Great Barrier Reef Marine Park Authority and Queensland Parks and Wildlife Service (G10/33405.1), with animal ethics committee approval from Southern Cross University.

3 Results

We recovered 95.2% of the complete STR panel across individuals through target capture—52,325 out of 54,978 possible STRs called—including all five markers overlapping capillary calls in all 22 samples. Across samples, a median 216 independent reads covered each marker, with between 13.0% and 18.9% of all reads per sample overlapping target STRs. After duplicate removal, this on-target proportion of reads equates to an effective 551-fold median enrichment of the target regions compared with whole genome shotgun sequencing data (range 393- to 807-fold). After SONiCS genotype calculation, we miscalled the capillary-validated genotype in only 1 instance—a 99.1% accuracy rate. SONiCS correctly rejected one additional genotype for failing the validation filters, yielding 108 genotypes for downstream analysis. The single miscalled locus was a heterozygous tetramer containing 10 and 11 repeats

erroneously called as a homozygote with 11 repeats (Table S1 and S3). Coverage of the miscalled locus was in the 11th percentile of all markers (48x), and visual inspection confirms that the majority of reads (n=37) supported the 11-repeat allele (Table S2). Thus this specific locus is a difficult-to-resolve case where manual calling would suggest a noisy homozygote—consistent with SONiCS—and where allelic dropout and capture biases in the presence of low coverage may have confounded the underlying genotype.

4 Discussion

Existing methods for stutter correction in genomic sequence data primarily involve training a noise model on a large haploid subset of experimental data, such as a human Y-chromosome (Gymrek *et al.*, 2012). This approach provides an effective and replicable genotype likelihood rescaling method for whole-genome STR datasets, but is not typically applicable to STR capture datasets: First, species without a chromosome-level reference genome or lacking large haploid chromosomes—most plants, for example—cannot make use of this haploid training approach. Indeed a major advantage of the BaitSTR method (Kistler *et al.*, 2017) used here for marker development is that no genome assembly is required, and this benefit is incompatible with the lobSTR training framework. Second, even given a large haploid genomic region, reduced representation experiments would need to devote a large proportion of probe sets to training markers that may not be applicable to broader research priorities, and the training process itself might be confounded by the extreme depth and variation in coverage typical in capture experiments. Finally, STR makeup and base composition in flanking regions (e.g. GC content) have the potential to influence capture efficiency, PCR uptake, and polymerase slippage during STR target capture. These variables would be difficult to unravel using a model-based noise correction method, but their influence is absorbed into the Monte Carlo approach using only weak uniform priors to constrain simulations.

For these reasons, we aimed to develop a stutter correction method that 1) can be used without a reference genome or an *a priori* noise model, and 2) is robust to the idiosyncratic

coverage variation and potential inter-locus biases of a target capture experiment. The statistical framework of SONiCS allows genotype selection in a strictly local context with only weak constraints on the parameters of STR fidelity, and therefore confers both independence from any reference genome and tolerance to highly variable genomic representation in a target panel. Genotype calls from SONiCS matched our truth-set of capillary STRs in all but one of the cases, experimentally demonstrating that the SONiCS approach is effective for de-noising sequence-based STRs containing PCR stutter.

Acknowledgements

We thank Jake Enk and Alison Devault (Arbor Biosciences) for helpful discussions about optimizing target capture. We thank Ani Manichaikul for helpful discussions about the method. We thank the Genetic Diversity Center-ETH Zurich for their support in data production. This work was supported by Southern Cross University (funds to DC) Sea World Research and Rescue Foundation (SWR/11/2016 to DC and MK) and Swiss Science Foundation (31003A_149956 to MK).

References

Baird,N.A. *et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS One*, **3**, e3376.

Behnel,S. *et al.* (2011) Cython: The Best of Both Worlds. *Comput. Sci. Eng.*, **13**, 31–39.

Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.

Elshire,R.J. *et al.* (2011) A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS One*, **6**, e19379.

Gnirke,A. *et al.* (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotech*, **27**, 182–189.

Gymrek,M. *et al.* (2012) lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.*, **22**, 1154–1162.

Kistler,L. *et al.* (2017) A massively parallel strategy for STR marker development, capture, and genotyping. *Nucleic Acids Res.*, **45**.

Schlötterer,C. and Tautz,D. (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.*, **20**, 211–215.